Lahore University of Management Sciences
Syed Babar Ali School of Science and Engineering
EE514 & CS535 - Machine Learning
Spring 2021 – March 19, 2021
Mid Exam

Time Allowed: 100 minutes **Total Points**: 100

Instructions

1. The exam is closed-book, closed-notes. Two hand-written A4-sized formula sheets (two-sided) are allowed. Calculators are also allowed.
2. Try to identify the easiest way to solve a problem.
3. Clearly outline all your steps. Solutions with inadequate justifications and/or steps may not receive full credit.
4. **The exam consists of TWO Parts.**
   (a) The first part is worth 60 pts, and is to be attempted on LMS. You will have 50 minutes to attempt the part 1 between 3:15 pm and 4:10 pm
   (b) The second part will be made available at 4 pm.
   (c) The second part consists of 6 problems worth a total of 55 points. We require to you to attempt questions with cumulative worth of 40 points. Do not over attempt; we will not grade the one with the highest marks.
   (d) We encourage you to solve the second part of the exam on A4 paper or electronic writing Pads. Use new sheet/page for each question and write sheet number on every sheet.
   (e) Your part must be submitted before 5:10 PM. No submissions will be accepted beyond this time.

EE514 & CS535 - Machine Learning

Mid Examination Spring 2021

Part - 1 (45 pts)

Problem 1. Select ALL (upto three) correct choices. Every incorrect answer would earn a penalty of 1 point but the total marks of any multiple choice question will not be less than zero.

(1) Which of the following is/are true about kNN algorithm?

(a) It is an instance based algorithm.

(b) Computational complexity to carry out prediction does not depend on the size of the training data.

(c) k-NN can be used for both classification and regression.

(d) We learn parameters of the algorithm during the learning stage.

(2) For kNN classifier, what tends to be correct about increasing the k in kNN algorithm?

(a) The decision boundary becomes smoother as k increases.

(b) The variance generally increases as k increases.

(c) As k increases, the bias tends to increase.

(d) The variance usually decreases as k increases.

(3) The dimensionality reduction is carried out to

- **Statement 1:** reduce in the computational complexity of **both** the training and prediction.
- **Statement 2:** remove the redundant features to improve the performance of classification or regression.

(a) Both statements are correct

(b) Both statements are incorrect

(c) Statement (1) is correct, statement (2) is incorrect

(d) Statement (2) is correct, statement (1) is incorrect

(4) For a data with d features, which of the following are true about forward search feature subset selection algorithm?

(a) It is an iterative algorithm that greedily adds the feature that most improves the prediction accuracy in each iteration.

(b) We are required to train $\mathcal{O}(2^d)$ models during the feature selection.

(c) It is a filter method.

(d) If we have few features relevant to the prediction, forward search feature subset selection is faster than backward search feature subset selection.

(5) Choose the correct answer(s) if we are getting high classification accuracy on the training data and low classification accuracy for the validation data?

(a) This is possibly due to overfitting.

(b) This is possibly due to underfitting.

(c) The training and testing data have been sampled from different distributions.

(d) None of the above.

(6) For a data with d dimensional feature space, we reconstruct or represent data using the first k principal components. Can we always (*exactly*) reconstruct any data point using k principal components?

(a) Yes if $k = d$

(b) Yes if $k < d$

(c) Yes if $d < k$

(d) No (always)

(7) Which of the following is/are true about dimensionality reduction using principal component analysis (PCA)?

(a) The features after dimensionality reduction are uncorrelated.

(b) PCA cannot be used to reduce the dimensionality if labels are not known for the data.

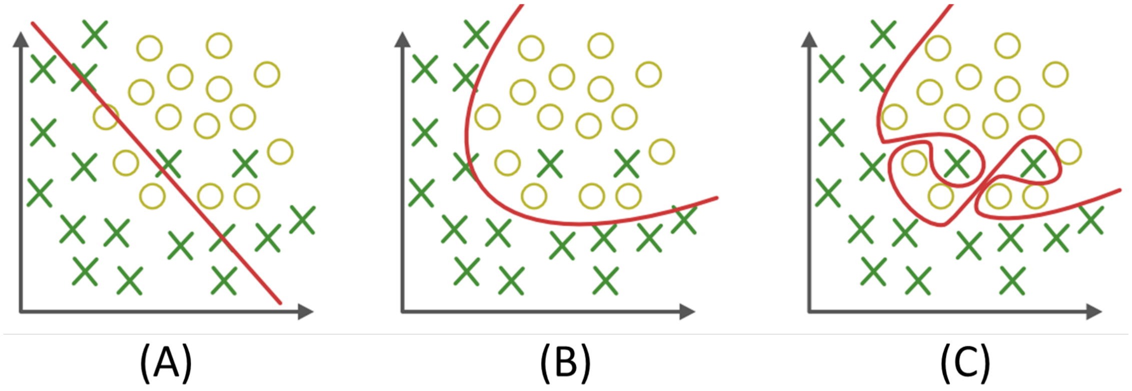
(c) PCA is a linear mapping.

(d) PCA maps the data along dimensions of maximum variances results in the new features.

(8) Overfitting in supervised learning can be avoided by

   (a) Introducing the penalty term in the loss function.
   (b) Minimizing error on training data.
   (c) Use train-validation split of the data. We expect to observe poor performance on validation data due to overfitting.
   (d) Use less number of points to ensure that model is not fitting to the noisy observations.

(9) For the following plots of the decision boundaries for three different classifiers A, B and C that have been trained on the same data, choose the correct statement(s).



(a) The classifier A has high robustness to noisy observations and is a poor fit.

(b) The classifier A has poor robustness to noisy observations and is a high fit.

(c) The classifier C has poor robustness to noisy observations and is a high fit.

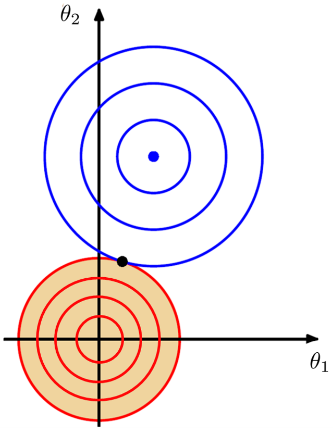
(d) The classifier C has high robustness to noisy observations and is a poor fit.

(10) The ratio of the volume of the unit ball shell of thickness of ϵ to the volume of the ball increases with the number of dimensions of the ball?

(a) True

(b) False

(11) For this question, consider the following figure depicting regularization during regression. We have a contour plot and and a constraint region for the representation of the error function (without regularization) and regularization term respectively. Choose the correct statement(s).



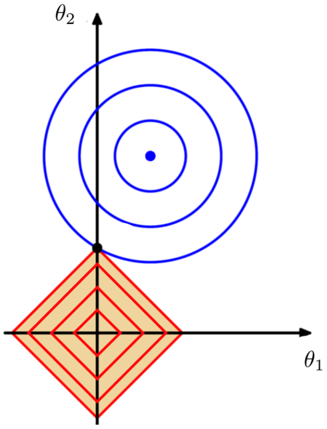
(a) This regularization is referred to as Lasso regression.

(b) This regularization yields sparse solution.

(c) We have used L_1 penalty as regularization term.

(d) None of the above

(12) For this question, consider the following figure depicting regularization during regression. We have a contour plot and and a constraint region for the representation of the error function (without regularization) and regularization term respectively. Choose the correct statement(s).



(a) This regularization is referred to as Lasso regression.

(b) This regularization yields sparse solution.

(c) We have used L_1 penalty as regularization term.

(d) None of the above

(13) We use cross validation for

(a) finding model parameters

(b) selection or tuning of hyperparameters

(c) avoiding or preventing underfitting

(d) avoiding or preventing overfitting

(14) A classifier trained on 100,000 training points with training accuracy of 99% gives accuracy of only 67% on testing data. Which of the following solutions can help us in improving the performance on the testing data?

(a) We should gather more data for training.

(b) We should set regularization parameter λ to 0.

(c) We should reduce the number of training points.

(d) We should use cross-validation to tune hyperparameters of the model.

(15) Consider a classifier for initial screening of Covid patients. True here refers to Covid positive. We do not want to diagnose any covid positive patient as healthy. Which of the following situations would you prefer for your classifier?

(a) FN $\gg$ FP

(b) FP $\gg$ FN

(c) FN $=$ FP$\times$ TP

(d) TN $\gg$ FP

(16) Which of the following quantities affect the trade-off between bias and variance?

(a) The regularization coefficient λ.

(b) The learning rate α in gradient descent

(c) The polynomial degree M in least-squares regression

(d) None of the above

(17) For polynomial regression using least-squares, which of the following is expected to go down at first but then go up with the increase in the polynomial degree?

(a) Variance

(b) Bias

(c) Training error

(d) Validation error

(18) Which of the following is/are always true about the ROC curve and the area under the ROC curve (AUC) for a binary classifier?

(a) ROC is obtained by varying the the threshold of the classifier.

(b) For a classifier with AUC=0.5, the performance is worse than the random guess (no power) classifier.

(c) The ROC curve increases monotonically with the false positive rate.

(d) The ROC curve explains the tradeoff between precision and specificity.

(19) Lasso can be interpreted as regularized linear regression that

(a) regularizes weights with the ℓ_1 norm .

(b) regularizes weights with the ℓ_2 norm.

(c) yields sparse solution.

(d) is computationally efficient than ridge regression.

(20) This question is related to kNN. Relative to 3NN, 1NN classifier has

(a) higher variance

(b) lower variance

(c) lower bias

(d) higher bias

(21) Consider building a spam classifier (binary: spam vs non-spam) for your mailbox. If we assume spam to be the positive class, which of the following would be more important to optimize? Note that we do not want to classify genuine emails as spam emails.

(a) Precision

(b) Recall

(c) Both Precision and Recall

(d) Accuacy

(22) Which of the following statements about Naïve Bayes is/are correct?

(a) Features are equally important.

(b) Features are dependent (statistically) of one another given the class label.

(c) Features are independent (statistically) of one another given the class label.

(d) Feature can be a categorical or numeric variable

(23) You are serving as a judge for the ML competition being organized by the school for freshman students. You are evaluating the group submissions with the following claims. Which ones would you consider rejecting?

(a) Our method gives a training error lower than the state-of-the-art method.

(b) Our method gives a test error lower than the state-of-the-art method when we choose the regularisation parameter λ that minimises the test error.)

(c) Our method gives a test error lower than the state-of-the-art method when we choose the regularisation parameter λ that minimises the cross-validation error.)

(d) Our method gives a cross-validation error lower than the state-of-the-art method when we choose the regularisation parameter λ that minimises the cross-validation error.)

EE514 & CS535 - Machine Learning

Mid Examination Spring 2021

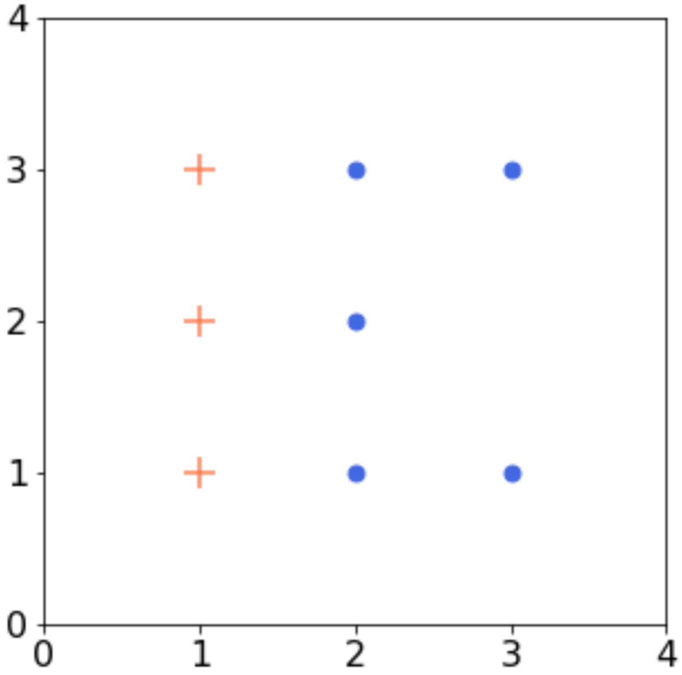
Part - 2 (40 pts)

Note:
There are six questions worth 55 points. We require to you to attempt questions with cumulative worth of 40 points. Do not over attempt; we will not grade the one with the highest marks.

Problem 1. (**5 pts**) For the following two dimensional labeled data, assign a label to the test point $(1.3; 4)$ using kNN for

(a) $k = 1$.

(b) $k = 3$.



Problem 2. (**5 pts**) In a rare collaboration, computer scientists, doctors and engineers developed tests for detecting Examophobia disease among the students. Research suggests that

- the probability of positive outcome for the student who is suffering from this disease is 0.8.
- the probability of positive outcome for the student who is not suffering from this disease is 0.01.
- 35% of the students suffer from Examophobia.

What is the probability that the student tested positive is suffering from Examophobia?

Problem 3. (**10 pts**)

Emirates Airlines have developed 2 different classifiers (A and B) for the prediction whether a flight originating from Dubai will arrive at its final destination on time or not. True or Positive here is 'On time' and it refers to the case when the flight that is no more than 5 minutes late as per schedule. The classifiers were tested on a data-set of 500 flights, and the results are as follows:

	Actual	
	On time	Late
Classifier A, predicted on time	131	155
Classifier A, predicted late	19	195
Classifier B, predicted on time	82	72
Classifier B, predicted late	68	278

Which is the preferable classifier in terms of F_1 score?

Problem 4. (**10 pts**) In Ridge regression, we minimize the following loss function

$$\mathcal{L}_{\text{reg}}(\mathbf{w}) = \frac{1}{2}\|(\mathbf{y} - \mathbf{X}\mathbf{w})\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

in order to find the model parameters $\mathbf{w} \in \mathbf{R}^d$. Here $\mathbf{X} \in \mathbf{R}^{n \times d}$ is the data matrix that is constructed using the features (inputs) in the training data, $\mathbf{y}$ represents the observations in the training data and $\lambda > 0$ is the regularization parameter.

(a) (**8 pts**) Derive the closed-form solution for the ridge regression problem, that is, find $\mathbf{w}$ that minimizes the loss function.

(b) (**2 pts**) How does the solution change as $\lambda \to 0$ and $\lambda \to \infty$?

Problem 5. (**10 pts**) We consider a data-set $D = \{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \ldots, (\mathbf{x_n}, y_n)\} \subseteq \mathcal{X}^d \times \mathcal{Y}$. We consider two dimensional feature space, that is $d = 2$, and each feature as a boolean (binary variable) following a Bernoulli distribution, that is,

$$P(x^{(1)} = 1|y = k) = \theta_{1k}, \quad P(x^{(1)} = 0|y = k) = 1 - \theta_{1k},$$

$$P(x^{(2)} = 1|y = k) = \theta_{2k}, \quad P(x^{(2)} = 0|y = k) = 1 - \theta_{2k},$$

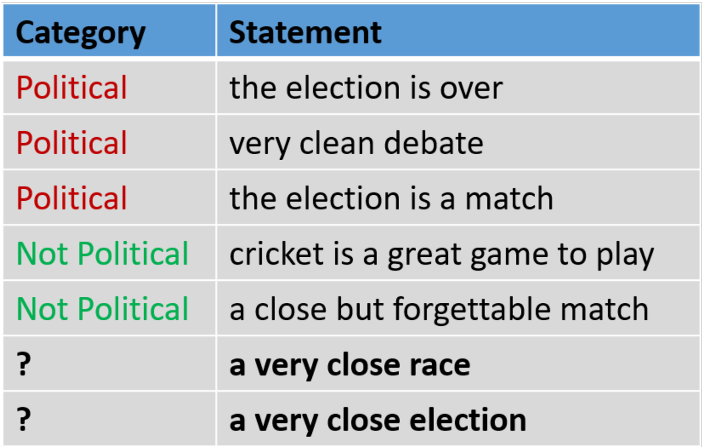
where $x^{(i)}$ denotes i-th feature of $\mathbf{x} = [x^{(1)}, x^{(2)}]$.

If we assume that all n samples are independent and identically distributed (iid), show that the maximum likelihood estimate of θ_{jk} is given by

$$\theta_{jk} = \frac{\text{count}(x^{(j)} = 1 \text{ and } y = k)}{\text{count}(y = k)}, \quad j \in \{1, 2\}.$$

Problem 6. (**15 pts**) This problem is related to Naïve Bayes classier for text classification that help us in classifying the given statement as 'Political' or 'Not political'.

Consider the following training and test data. We will build a Naïve Bayes classifier using bag of words approach to assign labels to the last two test statements.



Category	Statement
Political	the election is over
Political	very clean debate
Political	the election is a match
Not Political	cricket is a great game to play
Not Political	a close but forgettable match
?	**a very close race**
?	**a very close election**

(a) (**2 pts**) Create a vocabulary for the given training data.

(b) (**4 pts**) Develop a bag of words representation for each class using the training data.

(c) (**3 pts**) Do we need to use Laplace Smoothing for the given training and test data? Provide brief justification to support your answer.

(d) (**6 pts**) Classify the given test statements as 'Political' or 'Not Political' using the Naïve Bayes approach, that is, compute $P(\text{Political})P(\text{statement} \mid \text{Political})$ and $P(\text{Not Political})P(\text{statement} \mid \text{Not Political})$ for each test statement. Use add-1 smoothing for the computation of probabilities.